AMERICAN MARKETING ASSOCIATION MARKETINGPOWER.COM

MARKETING RESEARCH

IN THE MIX

Digital Sign of the Times: Context-aware computing

SPRING 2012

THE QUEST FOR BALANCE

A new, cost-effective technique tames scale use bias ... and nearly eliminates it

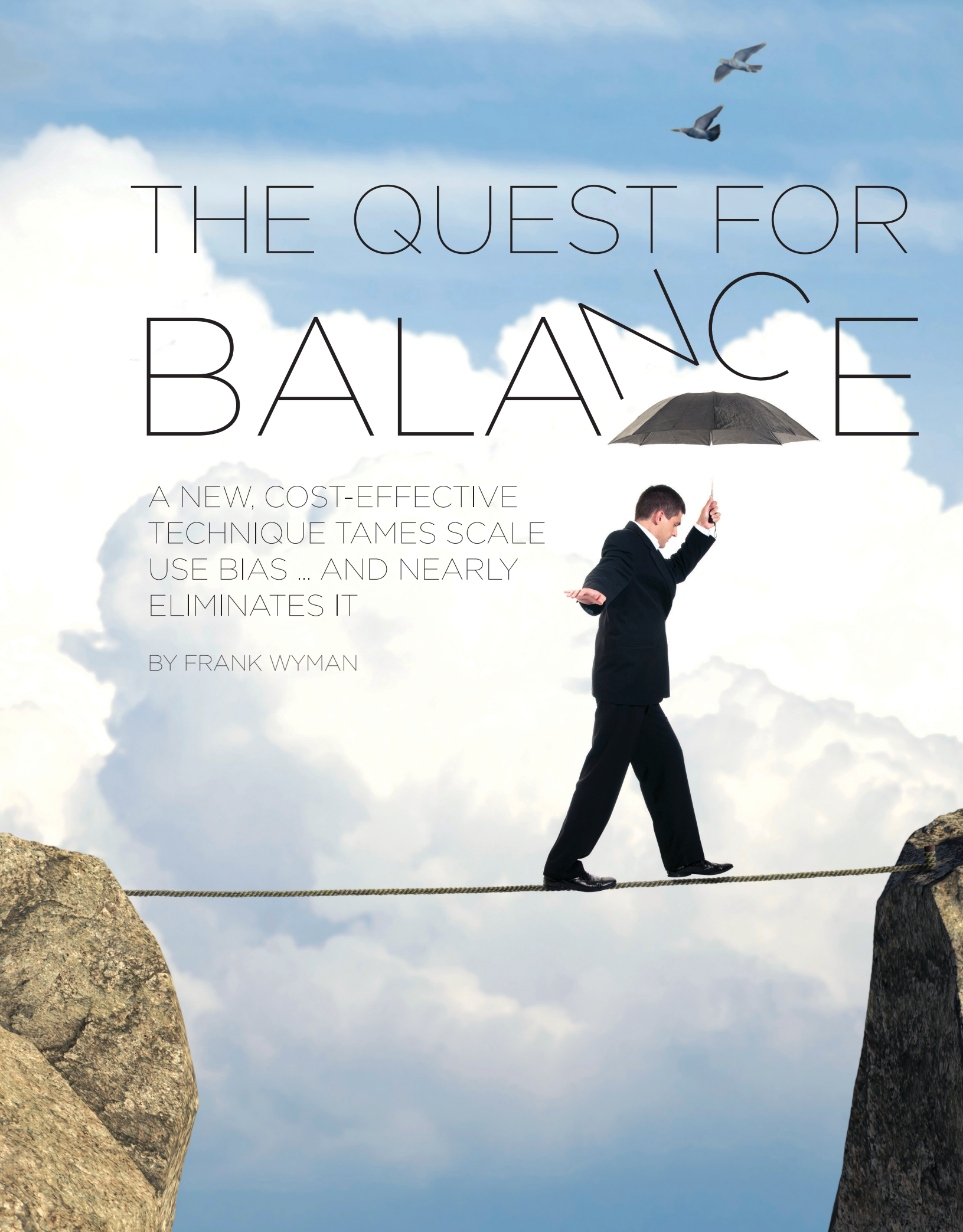
COMPLEX MARKETS: Bayes nets modeling untangles customer decisions / **NEUROMARKETING:** Reconciling science and speculation / **BRAND HALO:** The elephant in the room

75 AMERICAN MARKETING ASSOCIATION CELEBRATING 75 YEARS

THE QUEST FOR BALANCE

A NEW, COST-EFFECTIVE TECHNIQUE TAMES SCALE USE BIAS ... AND NEARLY ELIMINATES IT

BY FRANK WYMAN



Scale use bias occurs on a survey item whenever a respondent does not use the scale in the manner in which it was intended to be used. If two respondents have the exact same level of sentiment about something yet give two different responses, then the scale has failed to measure the sentiment in an unbiased fashion. Either one or both of those respondents has not used the scale as it was intended to be used, introducing bias into the data. In turn, that bias can greatly alter the outcomes of some types of data analysis and, ultimately, key study findings.

SOMETIMES SUCH BIAS occurs because the rating scale is poorly constructed, making it difficult for respondents to clearly discern how it is intended to be used. In only the most elementary type of analysis wherein only relative comparisons on basic descriptive statistics are sought, does it actually not matter that a scale is poorly constructed or, perhaps, purposefully constructed to be vague, as happens when using, for instance, a semantic differential scale. In other more advanced and/or absolute analyses, it becomes imperative that a scale be constructed so that it is completely clear in its intent and that each possible response point of the quantitative scale is also distinct and clear in meaning. Without proper framing of a quantitative rating scale, we cannot expect respondents to consistently use the scale in the correct manner. Vaguely defined intent, instructions and/or anchors (scale-point labels) of rating scales are a major invitation for scale use bias. (A list of best practices for creating rating scales that minimize scale use bias can be found at marcresearch.com/scaledevbestprac.html.)

However, the typical way in which scale use bias arises is arguably more related to respondents' personal or cultural proclivities toward using quantitative rating scales in certain systematic ways than to poorly constructed scales. Thus, no matter how well a scale has been constructed, a respondent's general manner of using quantitative scales can lead to responses that differ from what they should be. One respondent may simply never give a highest-scale-point rating. Another might tend to use only the higher end of the scale. And so forth.

Indeed, the list of response styles that can undermine the intended use of a given quantitative rating scale is long; and on that long list, unfortunately, we must also include that nefarious "style" that purposefully gives ill-conceived or even completely random responses. Whether because of a poorly constructed scale or because of a biasing response style, a response to a quantitative rating scale that differs from what it should be, given the scale's intent, is a response with bias, a bias that can be generally refered to as scale use bias.

You may well ask: So what? What's the big harm? Doesn't scale use bias just average out? The answer, surprisingly, is that, for a lot of typical survey research, scale use bias will do little to no harm. If, for example, a researcher wishes to know the relative degree of satisfaction consumers have for 10 different brands and ascertains this by comparing the means to a 1-5 satisfaction scale, then scale use bias poses no great threat. No matter how poorly the satisfaction scale is constructed nor how prevalent various response styles among survey participants are, the bias will essentially "average out" across the brands and have no untoward effect on the relative standings of the 10 brands. This "averaging out" assumes there is no systematic relationship between the degree of scale use bias and brand. The key

take-away from the research regarding relative brand satisfaction levels will not be altered due to the presence of scale use bias. It is only when a research objective requires results that must be interpreted in an absolute sense in terms of the scale's units and/or requires statistical data analysis involving correlation (e.g., regression analysis, factor analysis) or distance (e.g., cluster analysis, multidimensional scaling) between survey items that the ill effects of scale use bias will harm research conclusions.

In using a quantitative rating scale to answer a given research objective, two key questions should be asked:

1. Will answering the objective require absolute results wherein the units of the scale(s) in question are directly meaningful and will serve as an absolute reference during final interpretation, or will relative results suffice?

2. How will the responses to the scale(s) be analyzed?

a. Basic data tabulations: "Self-evident" results, such as percentages and means, will be directly tabulated from the responses to the scale(s).

b. Advanced statistical data analysis: Final results will be "derived" from responses to the scale(s) by way of statistical analysis involving correlation or distance.

Figure 1 provides a summary of the likely harm that scale use bias will have on final results given answers to these two questions. Because so much survey research tends to be of the type referenced by quadrant A of Figure 1, which is the only quadrant with a low risk of harm due to scale use bias, many survey research initiatives need not worry about scale use bias.

FIGURE 1.
THE RISK OF SCALE USE BIAS HARMING STUDY RESULTS

	Relative Results	**Absolute Results**
Basic Data Tabulations	A. Low risk of harm *(e.g., relative level of customer satisfaction for 10 brands)*	B. High risk of harm *(e.g., absolute level of voter approval of president on 10 attributes)*
Advanced Statistical Data Analysis	C. High risk of harm *(e.g. #1 via correlation analysis, the relative relationship (derived importance) of each of 10 performance attributes with overall satisfaction)* *(e.g.#2 via cluster analysis, the identification of consumer segments based on stated importance of each of 10 attributes)*	D. High risk of harm *(e.g., absolute degree of correlation (-1 to +1) between each of 10 shopping-style attributes and purchase intent)*

Another key aspect of a given study's use of a quantitative scale is the number of survey items (e.g., statements, attributes, brands) that are rated using the given scale. Three things are directly tied to that number:

- the options the researcher has for minimizing scale use bias during design;
- the researcher's ability to detect scale use bias; and
- the researcher's options for minimizing the effects of scale use bias once detected.

It is helpful to think in terms of three cases: Only one single item uses the scale, two to four items use the scale, and five or more items (i.e., a battery of items) use the scale. Generally speaking, when a quantitative scale is used on just one item, it is virtually impossible to detect how much of a response is true sentiment and how much is scale use bias. Consequently, there is no good way to adjust responses for scale use bias in the one-item case. Furthermore, data-gathering methods that constrain responses, which greatly protect against scale use bias, are not applicable. In the single-item case, then, it is crucial that best practices of scale construction be utilized so as to ensure the minimization of scale use bias by design.

On the other end of the spectrum, when a battery of approximately five or more items all use the same scale, it is easier to detect scale use bias. To do so, principal components analysis and k-means cluster analysis can be employed.

Responses to same-scaled items of a battery that are determined to be tainted with scale use bias are often adjusted through normalization, as in a "recentering" technique. The simplest such adjustment is to, for each respondent, subtract from each item score the respondent's mean score across all the battery items. This recentering process has been referred to as semi-ipsatizing and yields new scores that average to zero for all respondents. Since respondents may not only misuse the general level of a scale but also how they disperse responses across its range, a more thorough treatment is to fully ipsatize the data, which further adjusts the semi-ipsatized scores by dividing each by the respondent's unique standard deviation across all the battery items. This full ipsatization treatment thus adjusts (normalizes) the level and the dispersion of ratings for each respondent to be equal to the same level (mean=0) and dispersion (standard deviation=1). It is important to note that normalizations such as these can overadjust raw scores. Forcing the general response level to be absolutely equal (to zero) for all respondents may be unreasonable and inappropriate for some batteries, and such an adjustment may in fact cause its own form of unwanted bias. Furthermore, the loss of the ability to interpret results in terms of the original scale's units may be unduly restrictive for some research objectives. In short, attempts to ameliorate the effects of scale use bias through adjustments made after data collection may cause more harm than good.

A constrained-measurement method forces respondents to use each scale point some specified number (or specified range of number) of times. For instance, a ranking may be thought of as assigning as many scale points as there are items to rank wherein each point must be used precisely one time. With a battery of same-scaled items it may be appropriate to utilize a constrained-measurement method to completely or nearly completely eliminate scale use bias. It becomes important to think of batteries as being of two general types:

FIGURE 2.A:
BRAND Y PERFORMANCE RATINGS ON QUALITY AND PRICE

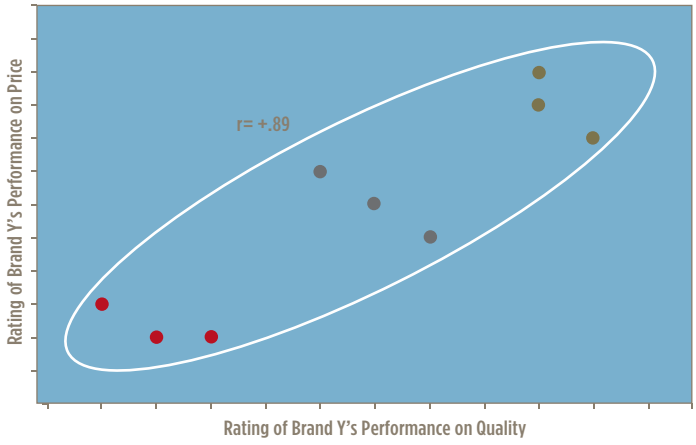
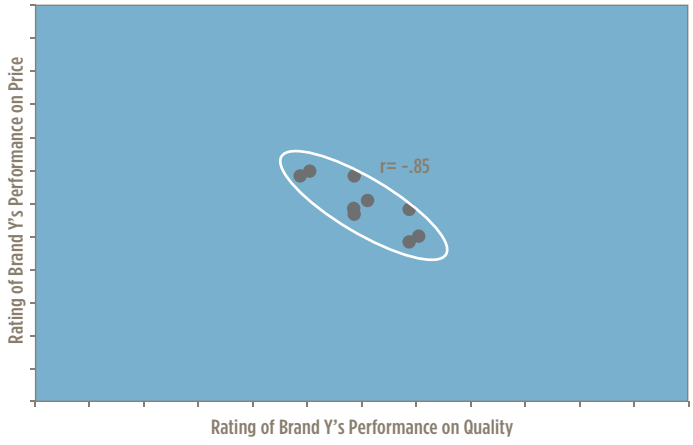


FIGURE 2.B:
BRAND Y PERFORMANCE RATINGS ON QUALITY AND PRICE



I. The mean response across the items of the battery is expected to be constant or nearly constant across respondents, or the objective at hand will not be undermined if either a constrained-measurement method (during data gathering) or a data-adjustment technique (during data analysis) is used to force those means to be equal or near equal.

II. The mean response across the items of the battery is not expected to be constant or near constant across respondents, nor does the objective at hand warrant forcing those means to be equal or near equal. Any measurement or adjustment to correct for scale use bias will also eliminate important inter-respondent differences in general response level, undermining study objectives.

For a type-I battery of items, there are options, over and above using best practices for scale construction, for protecting against and/or appropriately ameliorating the effects of scale use bias. For type-II batteries such options are not available or appropriate. Fortunately, it would seem most survey batteries are type-I batteries. For instance, it is common practice to ask how important each attribute in a battery of attributes is (to something such as purchasing a product in a given category). Such "stated importance" batteries can almost always be assumed to be self normalizing in the type-I way described here, or at least it can safely be assumed that if constrained to be normalized, the utility of the end results of the study would be enhanced. Even some types of agree-disagree or performance-level scales might be thought of as the type-I batteries for which it can be expected that respondents will respond generally (on average across all items of the battery) at the same level or that forcing such through a constrained-measurement method will only enhance end results.

For instance, if a battery of items is balanced and exhaustive with regards to all the aspects of what is being studied, it is quite reasonable to assume that each respondent will have a similar general response level (i.e., the distribution of the frequency of use of each scale point will be similar across respondents). In other cases, self-normalization should not be presumed, and by extension forced normalization through constrained-measurement or through a recentering type adjustment should not be pursued. The "type I vs. type II" aspect of a battery is critically important to deciding what the research might do regarding potential scale use bias. To help understand this somewhat imprecise aspect, a table of examples of type-I and type-II batteries has been developed and may be found at marcresearch.com/batterytypeexamples.html.

Finally, there is the aforementioned case of two to four items using the same quantitative rating scale. Generally speaking, the same points made regarding a battery of five or more same-scaled items applies to the case of two to four items, except that the two to four item list presents a bit of a gray area where all things do not work quite as well as they do with a battery of five or more items. While detecting scale use bias is a doable task, the detection will not be as clear-cut as with a longer list of items. And while one can normalize (recenter) two, three or four items in the same manners that can be done for a battery of five or more items, the effect of the normalization will produce less granular differences than when more items are present. Finally, while one can use certain constrained-measurement techniques such as ranking of the two to four items, a method such as FlexSort will not be able to be constructed for such a short list of items.

Two illustrations of how scale use bias manifests itself to cause extremely biased (wrong) results are now presented. Scale use bias typically causes an artificial increase in the correlation between variables (i.e., makes the correlation more positive than it truly is), thereby biasing any correlation-based analysis. Figures 2.A and 2.B provide a graphical example involving nine respondents' ratings on the perceived performance of Brand Y on Quality and Price. In the scattergram of Figure 2.A, assume the three respondents at the top right (green) misuse the scale by giving ratings that are generally higher than their true perceptions, assume the three respondents at the bottom left (red) misuse the scale by giving lower ratings than is true, and assume the three respondents in the middle use the scale correctly. Without knowing about the subgroups (which is typically the

FIGURE 3.A:
IMPORTANCE RATINGS ON SIX ATTRIBUTES

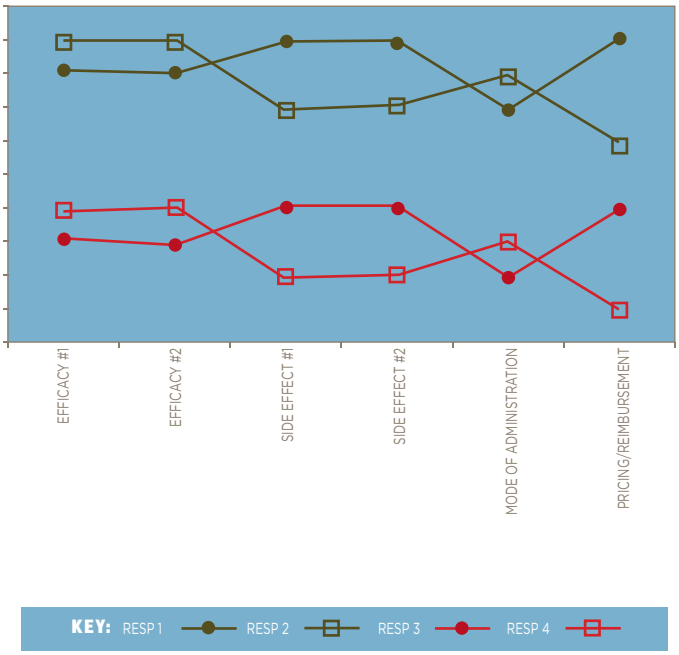
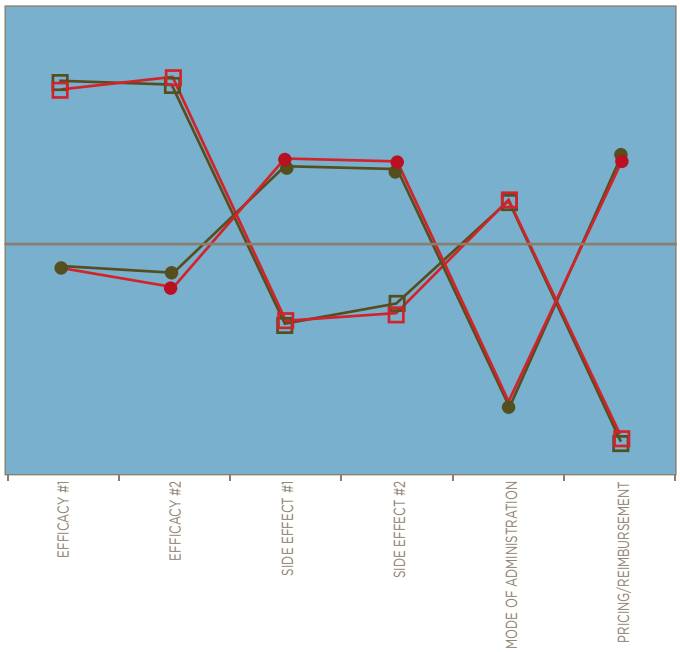


FIGURE 3.B:
IMPORTANCE RATINGS ON SIX ATTRIBUTES (RECENTERED)



case), we would calculate the correlation between Quality performance and Price performance as strongly positive (r = +.89): The better the quality, the better the price.

But Quality and Price performance, in actuality, should tend to have an inverse relationship, not a direct one, and so the above positive correlation does not make theoretical sense. Under closer inspection it can be seen that within each of the three response-style subgroups the relationship is the correct inverse one. And as Figure 2.B shows, had the two groups that misused the scale used it correctly, the correlation would indeed be negative (r = -.85). While overly simplified, this example typifies how markedly correlations computed on rating-scale data can artificially grow more positive in the presence of scale use bias. While sometimes mistaken for and discussed as high multicollinearity, artificial inflation of correlations due to scale use bias causes grave consequences in common higher-order analytics, such as regression analysis and factor analysis.

In Figure 2.B, the researcher's end goal might have been the estimation of the true absolute correlation between Price and Quality or alternatively the researcher might have been conducting a factor analysis among a large set of items in a battery (of which Price and Quality were but two). In either case, the effect of the scale use bias shown in Figure 2.B would definitely have been catastrophic to final results.

Scale use bias can also severely harm the results of statistical data analysis involving distances, the most prominent example of which is cluster analysis (aka interdependence type segmentation analysis). Since the essence of clustering involves distances, in the presence of scale use bias the analyst is more apt to identify segments that differ more in terms of how respondents have used the scale than in distinctive patterns across basis variables. Figures 3.A and 3.B provide a simplified example for four physician respondents' ratings of the relative importance of six prescription drug attributes. For this example assume the rating scale suffers from scale use bias such that the two green respondents rated generally higher than their true sentiment and the two red respondents rated lower than their true sentiment. A cluster analysis of this data would identify two clusters where one cluster contains the two green respondents and the other contains the two red respondents, basically a high-raters segment and a low-raters segment. But note how the patterns, across the six attributes, are nearly identical for respondents 1 and 3, as are the two patterns for respondents 2 and 4. In the absence of scale use bias, it is much more likely (and useful) that two clusters are formed by respondents 1 and 3 together and respondents 2 and 4 together.

In Figure 3.B note how a simple recentering (semi-ipsatization) of the importance scores for each respondent would yield a much more appropriate and useful cluster solution for the health care example, combining respondents 1 and 3 into a cluster that holds Side Effects and Pricing as most important, and combining respondents 2 and 4 into a second cluster that is primarily concerned with Efficacy. Without the recentering to adjust for scale use bias and as witnessed in Figure 3.A, clustering would have identified (actually misidentified) two segments, high raters and low raters, whose mean importance profile would have nearly the same identical, virtually flat pattern. Note that a constrained-measurement method could also have led to the correct responses presented in Figure 3.B. Sometimes researchers interpret some segments identified by cluster analysis of quantitative scale ratings as awkward, watered down or outright erroneous; perhaps these type segment results are due, at least in part, to a failure by the researchers to aptly protect against the effects of scale use bias. As alluded to already, one

constrained-measurement technique that can be employed to completely avoid scale use bias is to use rankings. To rank-order a list of items by definition sidesteps scale use bias completely since every respondent is forced to use the exact same "scale" (i.e., to give each and every item in the battery a single unique whole number between 1 and the number of items in the list). While there are advanced analytics available to deal with the ordinal form of data that results from rankings (e.g., the Spearman rank correlation coefficient for ordinal data) and to be sure many comparisons can be made just as well with mean rankings as with mean ratings, for many research studies rankings are simply not an appropriate form of measurement given the study objective(s).

Up until this point, only "stated" rating scales have been discussed. It is important to acknowledge that numerous "derived" measurement techniques have been devised that completely or nearly completely avoid scale use bias. For instance, the relative preferences and importances that are derived from conjoint analysis techniques are virtually devoid of scale use bias. Also, maximum difference scaling is a popular technique for deriving scores that possess little to no scale use bias. However, such derived techniques always cost more than simple stated scales because they require more up front design time as well as special analytics for data analysis on the back end. Because the added costs of derived measurement techniques can be quite substantial, for many research projects their use is not possible.

THE FLEXSORT TECHNIQUE

Recently M/A/R/C® Research introduced a new technique that nearly completely eliminates scale use bias yet is no more costly than conventional "stated" methods. FlexSort is a constrained-measurement technique that not only provides protection from scale use bias without requiring the extra costs of complex front-end designs and advanced back-end data analysis, but also provides respondents with a refreshing break from the conventional rating scale.

The FlexSort technique is an extension of the Q sort technique, which was part of the Q methodology (that used a special type of factor analysis called Q Factor Analysis) developed by William Stephenson in 1953. In the Q sort technique, m items are placed into a matrix with m entries or blocks, one place for each item. The matrix has typically been a symmetric triangle shape.

The major weakness of the conventional Q sort technique is that it forces a single distribution onto the response set, typically a symmetric triangular (pseudo-normal) distribution; all respondents must impose the exact same distribution over their ratings. While Q sorts have been used in market research for years, their use has mostly been limited to in-person modalities where actual paper tiles with item descriptions have been sorted onto paper matrices. To be sure, use of the Q sort technique has been quite limited; it is not a viable exercise for phone-based research, and research vendors have been slow to adopt the technique for online research.

M/A/R/C Research's new FlexSort technique has evolved the conventional Q sort technique in two important aspects. First,

Executive Summary

The data arising from quantitative rating scales is often influenced by scale use bias (also known as response style bias). This article first discusses some generalities regarding scale use bias including its end effects, its detection, how to reduce it during survey construction and fielding, and how to ameliorate it during data analysis. The FlexSort technique, a new, cost-effective technique for virtually eliminating scale use bias by way of a novel survey exercise, is then introduced.

M/A/R/C created a cutting-edge online programming logic that allows any type matrix to be quickly and inexpensively programmed. This programming logic has been developed over time so as to make the online clicking/dragging experience for respondents as easy, clear and enjoyable as possible. Underlying the logic is a proprietary set of rules governing the makeup of the matrix given each study's specific objectives.

The second important aspect of FlexSort is the incorporation of the functionality of allowing a flexible array of alternative distributions of response so that respondents are not bound to use the exact same static distribution as in the conventional Q sort. Thus, on a given study's battery of items, respondent Amy might truly have a flat (platykurtic) distribution while respondent Bob has a peaked (leptokurtic) distribution of response, and this is allowed to be revealed. The functionality of flexibility in response distribution allows all types and levels of skewness and kurtosis in response and provides the flexibility required to promote a feeling in respondents of not being overly restricted.

The FlexSort online exercise flows as follows:

1. Respondents are introduced to the rating task and instructed that all items are to be dragged and placed into the matrix of, for example, green and yellow blocks. The green blocks must be filled. The yellow blocks can be filled in any manner desired; there will be unused yellow blocks when finished.

2. Then respondent is shown the complete set of items to be rated (typically in randomized order).

3. Respondents are then presented each item in the set, one at a time, and asked to drag that item to the matrix column associated with the rating they wish to give the item. (Items "bottom stack" in each column.)

4. Once all items are in the matrix, respondents can make final adjustments and then, as long as all the green blocks are filled, hit a "finished" button to end the FlexSort exercise and proceed to the next section of the survey.

M/A/R/C Research has successfully used the FlexSort technique in many studies over the past year, with marked reductions in scale use bias and thus enhanced findings involving correlational and clustering analytics. The technique seems to work especially well for segmentation studies using cluster analysis of attribute importances as the basis of segment definitions. The results, which are always much more similar to the example presented in Figure 3.B than to the example presented in

Figure 3.A, rival those yielded in much more expensive studies where importances are derived from conjoint or other derived methods.

The FlexSort technique is ideal for a battery of items numbering 5-35 if the items are worded concisely and 5-20 with items whose wording is longer. FlexSort is also an appropriate way to minimize cross-cultural biases in ratings for studies spanning many different global regions and cultures. Returning to Figure 1, responses yielded from the FlexSort method are completely appropriate as input into the "relative" analysis types A and C, mostly appropriate as input into analysis type D, and almost never appropriate as input into analysis type B.

FlexSort is not only useful for importance ratings. It can work equally well for agree-disagree, performance and other types of scales as long as the battery of items is a type-I battery, as discussed earlier. Even for a scale with a battery of items that is not strongly/absolutely a type-I battery, the FlexSort method can work very well by relaxing the degree of constraint on the respondents by way of including more "yellow blocks" (flexibility) than usual.

Scale use bias is a pervasive problem in survey research and deserves the attention of all researchers, especially those who conduct advanced statistical analysis involving correlation and distance. For obtaining quantitative ratings on items of a same-scaled battery, FlexSort offers the right balance of flexibility and rigidity to nearly eliminate scale use bias while allowing for respondent heterogeneity and not over-restricting each respondent's true sentiments. **MR**

FRANK WYMAN is vice president of Advanced Analytics at Dallas-based M/A/R/C® Research. He can be reached at Frank.Wyman@marcresearch.com or (864) 938-0282.